Email Archiving Systems Interoperability

Harvard Library Report
July 2016

Prepared by Joel Simpson

HARVARD LIBRARY

Prepared by Joel Simpson, Artefactual Systems, Inc.

Reviewed by Wendy Marcus Gogel, Harvard Library and Grainne Reilly, Library Technology Services, Harvard University

Citation:
Simpson, Joel. 2016. Email Archiving Systems Interoperability. Harvard Library Report. http://nrs.harvard.edu/urn-3:HUL.InstRepos:28682572.

Table of Contents

Executive Summary

Earlier this year, Harvard Library convened the Harvard EAST (Email Archiving Stewardship Tools) workshop to foster the expanding email archiving community, share best practices and identify directions for future work.

One of the main conclusions of the workshop was that there is no standard workflow that can be uniformly applied in every situation, but that all archives have similar functional needs for email archiving, and that given the need for flexibility, current processes could be improved by using the unique strengths of different tools together.

Harvard Library engaged Artefactual Systems Inc. to better understand how the tools can exchange data today and carry out analysis to identify opportunities for the community to further support comprehensive preservation workflows for email.

Community members have been invited to contribute to an **Email Tools Data Sharing Framework**. The intention is to provide a high level view of how email content or metadata can be input or output to each of the different tools, using a common framework to support comparison and analysis. This work is ongoing, but enough detail has been collected to enable analysis and identification of some clear opportunities for improving the interoperability of these tools.

A set of "requirements for interoperability" were identified to set out the different aspects or concerns involved in using multiple tools in an email archiving, processing or preservation workflow. Analysis was carried out to understand how each of the tools supports these different requirements. Key findings were then identified in each of these areas.

Finally, a set of 7 draft recommendations has been proposed for the wider community to consider. These are high level recommendations without detailed nexts steps or any suggestion for priority. We feel they are useful in decomposing this complex problem space into discrete and well-defined opportunities that will be easier to tackle in a fast changing environment.

Background and Context

Earlier this year, Harvard Library convened the Harvard EAST (Email Archiving Stewardship Tools) workshop to foster the expanding email archiving community, share best practices and identify directions for future work. The workshop involved stakeholders from different institutions, including subject matter experts, users and developers of several email archiving or preservation tools.

The workshop concluded that the community is very interested in working together to solve shared problems. Several directions for future work were identified, including "the need for an exchange standard that enables interoperable ways to extract, package and transfer data between tools". This conclusion was based on the consensus that there is no one uniform workflow for email archiving, but that current processes could be improved if archives were able to harness the unique strengths of each tool selectively (using only the functionality needed in whatever order is needed).

Harvard Library Preservation Services engaged Artefactual Systems Inc. to carry out a short consulting project to build on these findings and identify opportunities for the community to further support comprehensive preservation workflows for email.

Project Objectives

The goals of this consulting project are to:

1. identify gaps or opportunities to improve the interoperability of the numerous email tools by showing the type, format and structure of data which can be input or output from each tool
2. inform email stewards about the options and considerations involved in defining email archiving workflows using multiple tools

This project has not attempted to provide a functional description or comparison of the various tools under consideration. A very brief overview of the tools, with links for further detailed information available from the providers, is provided below in section 3. A useful comparison of Email Archiving tools (including many not considered in this project) can be found at the Lifecycle Tools for Archival Email Chart:https://docs.google.com/spreadsheets/d/1V1N22xnr5e0EbDlZWx58bjYO6rkrMrYH9wGX9-CK8c4/edit#gid=986222267.

Project Approach

This project is producing two deliverables to meet the objectives defined above.

The first deliverable is an **Email Tools Data Sharing Framework** that sets out the content objects (i.e. email) and metadata that each email or preservation tool can input or output. Representatives from each tool provider were asked to complete the descriptions of these inputs and outputs using a generic framework (with associated glossary) to enable common understanding of terms and make comparison between tools easier.

A more detailed description and assessment of the tool is provided below in section 2.

The second deliverable of this project is this **Consulting Report** which

1. assesses the completion and usefulness of the Email Tools Data Sharing Framework
2. proposes a generic set of requirements for interoperability to use as an analysis framework
3. analyzes / summarizes how each tool satisfies those requirements for interoperability
4. sets out several recommendations for improving interoperability of the tools and further establishing best practices for the community

Please note that throughout this report when we refer to 'digital objects' we mean any type of digital objects, including emails themselves, related content like attachments, or any associated metadata. We use 'data' interchangeably with 'digital objects' simply because it is shorter. (We have not seen the need to distinguish these concepts with more precise definitions.)

Project Results

1. Assessment of the Email Tools Data Sharing Framework

1.1. About the Email Tools Data Sharing Framework

The email tools data sharing framework includes information on 6 different email or preservation tools. The intention is to provide a high level view of how email content or metadata can be input or output to each of the different tools.

The framework is set out in a spreadsheet, with one sheet to describe inputs and another to describe outputs. Each sheet is organized to first describe the actual (or "physical") data objects (or input/output mechanisms, as in some cases they are programmatic), followed by a description of the kinds of data or metadata found in those objects.

Separate rows distinguish between the level of obligation demanded to be able to use each tool:

- mandatory content or data (system will not accept or work properly without this)
- useful content or data (is optional, but enables functionality within the system - e.g. a sensitivity flag that can be used when filtering)
- additional content or data (can be consumed, but is not used in any way by consuming system -- e.g. attachments are included in MBOX, but the particular system may not allow users to do anything with them)

The goal is to describe in each of these columns:

- the type or extent of data provided (e.g. specific fields used as reference IDs, or a more general description such as 'preservation events')
- format of data (is a 'local' schema defined, or is a standard schema used, such as PREMIS)
- location / structure of data (where in the input / output is this information -- e.g. PREMIS events are recorded in METS.xml file; folder information stored in pathname in MBOX etc.)

In some cases this information needs to be broken down into different levels of granularity, for instance to indicate information stored at individual email level vs. collection level.

1.2. Assessment of the Email Tools Data Sharing Framework

At the time of this writing, completion of the spreadsheet is in progress. We invite comments or thoughts from all participants on:

- ability to complete the spreadsheet consistently (or key differences in interpretation)
- anything learned while filling it in
- whether it is complete enough, or needs further work; wishlist additions / amendments (e.g. suggestions for adding more detail)
- initial views on value of the exercise
- intent to use the tool moving forward

Data gathering work is ongoing and will be refined as needed by the community to support their collaborative efforts to improve these tools and establish best practices for email archiving and preservation.

Initial feedback and observations from Artefactual:

- It is interesting to see this particular perspective from the different tools, and enables interesting analysis of similarities and differences (which will be explored further in the rest of this report).
- The spreadsheet emphasizes two dimensions (data types in columns and systems in rows), but there are in fact numerous dimensions of interest (including granularity of grouping of data, levels of obligation, type of data vs. formats or standards employed, etc.). This makes fitting in all of the relevant information a challenge.
- Given the space, it does not seem possible to include enough detailed information for this to be a very hands on 'how to' tool -- but it may well be a useful analytic or decision support tool, to determine if there is enough compatibility between a particular selection of tools for a desired workflow.

2. Analysis Framework: Requirements for Interoperability

The data sharing framework is primarily focused on the inputs and outputs of each of the tools under consideration. Given the broader intent to enable email stewards to determine whether and how they might craft workflows using multiple tools, this report proposes a set of generic 'requirements for interoperability'. This provides a more holistic view of the different aspects of using multiple tools that operate together to enable a comprehensive workflow for email processing or preservation.

These requirements are more an analytical framework than a concrete set of requirements. They are focused on the level of business processes and workflows, and do not represent a particular effort to elicit requirements from end users.

The requirements and their rationale are described below. In the following section, each of the 6 tools is assessed against each requirement. This allows us to compare similarities and differences in specific areas of concern and use this as the basis for recommendations for future work later in the report.

2.1. Support for data transmission

The most basic requirement for a workflow that uses multiple tools working on a common set of data is to enable those tools to access that data.

This functionality can be provided in many forms; user interfaces for selection of data for ingest from a particular location; automated jobs that ingest data; direct system to system connectivity; or published APIs. The goal here is to simply articulate how each system supports this, rather than to judge one method over another. This will allow us to see which tools can share data (and how), at a physical level, with other tools.

2.2. Support for standard data formats

Once we have determined a particular tool can access a set of data physically, we need to ensure it can interpret and process that data. At a minimum, the data format must be 'standard' between the tools being considered.

It is well established in the preservation community that open, non-proprietary and widely used standards are preferable for preservation formats. While not all data to be exchanged needs to be (or even can be) in a preservation format, the same principles will improve the odds that any particular tool will be interoperable with others.

Support for standard data formats applies to email content, metadata and the packaging of both email and metadata.

2.3. Support for appropriate scope of exchangeable data

Email content and metadata can exist or be grouped at various levels of granularity. Different processing tools may accept data with an entirely arbitrary definition of scope (using a generic term such as a 'transfer' or 'packet'), or they may require data or metadata to conform to a specific definition (such as clearly grouping data by 'account').

Scope of data also refers to the type and extent of data in any particular data set. For example, Archivematica has functionality to verify hashes / checksums; if checksums have been created in another tool (e.g. BitCurator), then ideally Archivematica should allow checksums to be imported so that verification can occur on those checksums, not just on checksums created by Archivematica. This concept is clearly tied closely with the level of granularity - a checksum may be made for a folder or collection of emails, or it may be created at the individual email level.

Email stewards will need to understand what scope of data is required or possible using any particular tool. Similarly any decision to use a particular data standard needs to consider the scope of data that format allows for or requires.

2.4. Ability to track processing history and provenance

The ability to establish and maintain the provenance (including processing history) of content is a well understood requirement in the archival and preservation communities. While this may not be a requirement for everyone looking to process emails, it is a fundamental requirement for the core user groups of many of the 6 tools we are evaluating.

Email stewards who do need to record and capture provenance will generally need a mechanism to do this whenever they are processing, creating or changing data. This means that either the tools they use for processing need to capture processing history directly, or they need some ability to track processing history manually and store it appropriately.

2.5. Support for maintaining the identity and integrity of data

As data is moved, migrated or processed by different tools, email stewards need to be able to ensure that the identity and integrity of the data they are processing is not compromised.

Maintaining the identity of the dataset depends in large part upon using identifiers to link it to its descriptive and administrative metadata, and ensuring that this link cannot be broken. Most tools generate unique identifiers, but these are usually local (assigned, stored and maintained within the tool itself). External identifiers may be supported, either informally (e.g. by recording an accession number as part of a directory structure or filename) or more formally (as in having a field with a declared data type that aligns to the identifier used by another system). Some systems also support identifiers that refer explicitly to external resources or authorities (a concept underpinning linked data).

Maintaining the integrity of digital objects is often achieved using hashes or checksums, with regular verification, to ensure that the content of the ingested data has not been altered over time. The hashes or checksums can be assigned to both the original ingested content and to any normalized or otherwise modified versions that may be generated from that content. Hashes or checksums may also be assigned to associated metadata.

Another common practice to safeguard the integrity of data is to package content and metadata 'together' for transfer, reducing the risk of corruption or loss (i.e. links between the two breaking at some point).

2.6. System access and documentation to support interoperability

A basic requirement is the ability to access and use the software, both technically and with appropriate permissions or licensing.

All of the capabilities mentioned above are less useful in practice if knowledge to use them is not captured well. Technical and user documentation, training materials and training resources (i.e. trainers for hire) all add to the ability to use the tool as part of an integrated workflow. The starting minimum is documentation on how to use the tool at all. Ideally a knowledge base would address the exchange of

data, interoperability with other systems and any license requirements.

3. Analysis of Tools using the Requirements for Interoperability Framework

3.1. Archivematica

Archivematica is an integrated suite of open-source software tools that allows users to process digital objects from ingest to access and to implement preservation plans. Users monitor and control ingest and preservation micro-services via a web-based dashboard. Archivematica uses METS, PREMIS, Dublin Core, the Library of Congress BagIt specification and other recognized standards to generate Archival Information Packages (AIPs) for storage in external repositories.

Requirement	Supporting Functionality	Observations
Support for data transmission	Digital objects need to reside in a locally accessible filesystem for ingest. Archivematica is provided with an accompanying application called Storage Services that can be used to configure access to sources of data for ingest. There is an API to assign accession numbers, but no direct support for moving data across hardware, networks etc.	There are numerous external tools available for moving data.
Support for standard formats	Any digital object can be ingested, so any email format can be processed with core functionality. Email input in MBOX format can be processed using additional functionality (extracting attachments and metadata). Email input in maildir can be normalized and output as MBOX. The BagIt file packaging standard is supported for input and output. Metadata input in csv or json formats can be processed. Additional metadata (in other formats) can be included but not processed. Metadata outputs are well supported by widely adopted standards (METS, Dublin Core, PREMIS, Bag)	No support to normalize to EML format (widely used email format).
Support for appropriate scope of data	Transfer, Submission, Archival and Dissemination packages can be structured and described using any definition the user chooses. For example, an email account or accounts can be ingested as one or more SIPs, and multiple SIPs can be combined into one or more AIPs. Some key metadata, such as rights metadata, can only be input or assigned during processing at the package level.	Provides complete flexibility but no native support for common email groupings (e.g. account, folder etc.) Rights metadata can't be assigned to individual emails, so users would have to manually structure inputs and outputs to reflect different rights (e.g. create one AIP or DIP for restricted emails, and one for

		non-restricted emails).
Ability to track processing history and provenance	Provides extensive functionality to track processing history and record using PREMIS Processing history from external sources could "travel with" any data sets, but currently no ability to merge or consolidate processing history from multiple systems.	Email stewards could create manual processes to maintain multiple processing history files.
Support for maintaining the identity and integrity of data	Archivematica assigns UUIDs to all ingested objects and uses the UUIDs and ID attributes in the METS files to maintain links between digital objects and their metadata. Archivematica also supports a wide range of external metadata, so there are several ways external identifiers (i.e. from other tools) can be maintained. However there is no direct support for typed / declared external identifiers (e.g. automatically adding identifiers when importing from an external system). Fixity verification is supported using both internally or externally created hashes.	Email stewards could create manual processes for aligning and maintaining referential integrity across systems (but may need to plan this - e.g. aligning package structure to external identification systems)
System Access and Documentation	Documentation available, community support website / groups, as well as for hire services for consultancy, training etc. Source code and technical info available on GitHub. Documentation can be quite technical.	

3.2. ArchivesSpace

ArchivesSpace is an open source, web application for managing archives information. The application is designed to support core functions in archives administration such as accessioning; description and arrangement of processed materials including analog, hybrid, and born-digital content; management of authorities (agents and subjects) and rights; and reference service. The application supports collection management through collection management records, tracking of events, and a growing number of administrative reports. The application also functions as a metadata authoring tool, enabling the generation of EAD, MARCXML, MODS, Dublin Core, and METS formatted data.

(summary taken from: https://archivesspace.atlassian.net/wiki/display/ADC/ArchivesSpace)

ArchivesSpace is not a digital asset or document management system and cannot manage digital files or digitization workflows. The digital objects module can be used to describe digital objects and link to digital files stored elsewhere. The metadata created can be exported to other systems as MODS, METS, or Dublin Core or made publicly accessible through the built-in public interface, though the viewers in

the public interface are more limited in their functionality than those of a digital asset management system or digital repository.

(detail on digital objects taken from FAQ: http://www.archivesspace.org/faq)

Requirement	Supporting Functionality	Observations
Support for data transmission	ArchivesSpace does not provide a means of moving or storing email content. Metadata can be exchanged as files or through a set of APIs.	
Support for standard formats	ArchivesSpace supports a range of well established standards for describing archival records - EAD, MARCXML, MODS, Dublin Core, and METS formatted data. ArchivesSpace does not support functionality or processing of email content (i.e. normalisation, search or identification of authorities etc.)	
Support for appropriate scope of data	ArchivesSpace provides functionality for describing the arrangement and relationships of digital objects. It does not support email specific concepts directly (e.g. the notion of an email account)	It could be useful to establish conventions or best practices for describing email accounts and their potential relationships to collections, agents etc.
Ability to track processing history and provenance		
Support for maintaining the identity and integrity of data	Support for identifiers and integrity internally within a repository. The system supports structured capture of agents and subjects which will improve consistency and accuracy of description	
System Access and Documentation	ArchivesSpace is an open source project with considerable documentation available. It is supported by the Lyrasis organisation with full time staff who are developers and subject matter experts.	

3.3. BitCurator

The BitCurator Environment is built on a stack of free and open source digital forensics tools and associated software libraries, modified and packaged for increased accessibility and functionality for

collecting institutions. The BitCurator software is freely distributed under an open source license. It can be installed as a Linux environment; run as a virtual machine on top of most contemporary operating systems; or run as individual software tools, packages, support scripts, and documentation.

Key features of BitCurator include:

- Pre-imaging data triage
- Forensic disk imaging
- File system analysis and reporting
- Identification of private and individually identifying information
- Export of technical and other metadata

(summary taken from: http://www.bitcurator.net/bitcurator/)

Requirement	Supporting Functionality	Observations
Support for data transmission	BitCurator does provide support for migrating data without altering it in any way, starting with the concept of creating forensic images before further transmitting or processing data. Uniquely among the tools considered here, BitCurator provides software write-blocking functionality to ensure the integrity of source objects.	As this is an area not well supported by other tools, it could use some elaboration / detail.
Support for standard formats	Supports DFXML (Digital Forensics XML) that enables the exchange of structured forensic information. BitCurator generates PREMIS metadata when the user runs several of its core data forensics tools, providing a record of key processing events. Provides some processing support for email - e.g. using readpst to convert PST email objects into MBOX. Also supports BAG format for output.	
Support for appropriate scope of data	The BitCurator environment includes numerous applications to be used for different purposes, to be run against individual items or collections of terms. One of the most commonly used tools is bulk_extractor, which can be used to identify potentially sensitive information on disks, disk images or directories. Other core tools, including fiwalk and other specialized reporting tools, are designed to be run against entire disk images. When run against a disk or disk image, bulk_extractor reports on the location of patterns based a byte off-set onto the disk. Other reporting tools, including fiwak, generate metadata based on the filesystem (files and folders). In the case of email, the files would be likely in formats such as .pst or mbox. Those wishing to	

	generate metadata associated with specific messages within those container files could use readpst and pipe its output to other command-line tools. BitCurator is primarily concerned with identification and description of digital objects rather than arrangement.	
Ability to track processing history and provenance	BitCurator generates PREMIS metadata when the user runs several of its core data forensics tools, providing a record of key processing events.	Email stewards could create manual processes to maintain multiple processing history files.
Support for maintaining the identity and integrity of data	BitCurator provides support for indexing, characterizing and uniquely identifying all content on a disk or disk image. Bitcurator supports creation and validation of hashes / checksums.	
System Access and Documentation	BitCurator is an open source project with considerable documentation available.	

3.4. DArcMail

DArcMail (for Digital Archive Mail System) was created by the Smithsonian Institution Archives. DArcMail provides normalization, item level and bulk processing, intellectual arrangement, search capability, packaging and access functionality for email.

Requirement	Supporting Functionality	Observations
Support for data transmission	Digital objects need to reside in an accessible filesystem for ingest.	
Support for standard formats	Email input requires MBOX as the original format or as an interim normalization format. Email input in MBOX format can be processed with all core functionality including exporting preserved emails, email collections or email accounts in the EMail Account XML (EMA). EMA is a comprehensive XML schema designed for RFC 5322 compliant preservation purposes applied to the full range of email objects, i.e., single message to whole email account. All elements of the original email is retained in the preservation EMA XML output. User-defined subsets of email messages can be created and exported in MBOX or EMA XML formats.	No support to normalize to EML. The EMA XML schema is not widely adopted. It is fully implemented in two other email archiving tools, or in limited fashion in a couple other applications.

Support for appropriate scope of data	DArcMail allows users to interact with emails on an individual, group or account basis. Complex searching, filtering and message thread tracking. Attachments can be searched, viewed and separated from email.	
Ability to track processing history and provenance	The DArcMail tool is designed to be used for initial appraisal and then for preservation (AIP) and access (DIP). It natively retains the logical arrangement of the original account in both the AIP and DIP packages. Its flexibility allows for creation of custom subsets of email for creation of specialized AIPs and DIPs.	Transfer and accessioning of email digital objects occur outside of the DArcMail workflow. Non-technical metadata such as rights metadata must be captured and maintained in a separate system or manually.
Support for maintaining the identity and integrity of data	DArcMail maintains all UIDs present in the original emails. It generates SHA-1 checksums for each message and for email accounts as a whole which are embedded in the EMA preservation format. DArcMail also produces external metadata including the checksum for each message preserved.	The internal message and account checksums are retained even if the preserved email account is moved to from one repository to another.
System Access and Documentation	DArcMail is not currently available outside of the Smithsonian. Limited documentation is publicly available. The Smithsonian intends to release it as open source when time / effort allows.	Making the tool publicly available is a precondition for any other community users.

3.5. Electronic Archiving System (EAS)

Harvard developed the EAS tool to enable archival processing of email messages and attachments and automate the process of making deposits to Harvard's preservation repository. Key features include:

- Normalization to EML -- an open standard for preservation (an extension of IMF RFC 5322) -- for long term preservation.
- Summary views of the metadata associated with email or attachments within a result set.
- Batch and item level processing options for archivists.
- Long term preservation of email and attachments in a secure environment approved for sensitive data is supported by automated packaging and transfer to the preservation repository – Digital Repository Service (DRS).
- Capture of essential rights management information using PREMIS.

- Capture of significant events tracking to document deletions of email and attachments and format transformations such as the conversion of the native mail format to EML.

(feature list taken from: http://hul.harvard.edu/ois/systems/eas/)

Requirement	Supporting Functionality	Observations
Support for data transmission	Data need to be moved to a 'dropbox' (directory space in Harvard systems). EAS documentation describes how to use a secure FTP client to move the data but this is not part of the EAS solution.	There are numerous external tools available for moving data.
Support for standard formats	Email content can be input in MBOX or PST format (which covers the majority of email client standards for output of email). Attachment objects of any type (e.g. .ppt, .doc) can be embedded in the emails or provided separately. It is not possible to input metadata (beyond that contained directly in MBOX/PST or attachment formats). Email is output to EML format, with attachments extracted. Overall metadata is captured and output using well established standard formats (e.g. METS and MODS) and both rights and processing history are captured in PREMIS. Some reference metadata is in local format defined by Harvard (for packets, collections etc.), as is metadata relating to security (access) and sensitivity (using locally defined 'flags').	Email content formats well supported. While EML format for output is a well established standard it is not accepted by all other tools for input. Security and sensitivity metadata could potentially be captured using more widely used standard. Referencing metadata geared towards Harvard integration with DRS system. May not be any need to standardize this, but support for external IDs would enable better interoperability with other tools.
Support for appropriate scope of data	Submission packets can be structured and described using any definition the user chooses. It is not possible to input additional metadata or content beyond email / attachments. Processing work can be completed at individual item level (email or attachment) or at various levels of grouping (folder, collection etc.). Additional groupings can be added (collections or series). Outputs will always contain the same packet structure as the associated input. Output contains normalized / processed content; does NOT contain original input files (i.e. in MBOX or PST format)	Provides support for grouping (in collections etc.) Inability to input additional metadata or content suggests this tool may work best at 'start' of a workflow. Stewards will need to think through manual processes for managing metadata created using other tools.
Ability to track processing history and provenance	Provides functionality to track processing history and record using PREMIS. No ability to merge processing history with that from other tools.	Email stewards could create manual processes to maintain multiple processing history files.

Support for maintaining the identity and integrity of data	Identifiers are internal (e.g. EAS message ID) or local to Harvard (e.g. DRS codes are for Harvard repository). Integration with 'Wordshack' application ensures some descriptive or identification information is based on controlled vocabularies used in Harvard (i.e. also integrated with Harvard DRS repository). This improves consistency in use of admin categories and topics, and improves identification quality for persons or organisations.	Support for external referencing systems would better enable multi-tool workflows. Use of controlled vocabularies limited to Harvard currently - could be several approaches to extend this - e.g. publishing those vocabularies as open data, or enabling use / integration of other (e.g. linked open data) vocabularies as alternatives
System Access and Documentation	User documentation available and support for Harvard users. System is not currently available beyond Harvard users.	A project has been proposed to release system as Open Source project; but some technical work required to make ready for more generic use.

3.6. ePADD

ePADD is a software package developed by Stanford University's Special Collections and University Archives that supports archival processes around the appraisal, ingest, processing, discovery, and delivery of email archives. The user guide (https://docs.google.com/document/d/1joUmI8yZEOnFzuWaVN1A5gAEA8UawC-UnKycdcuG5Xc/edit#) provides the following description of the major modules in the system:

> **Appraisal:** Allows donors, dealers, and curators to easily gather and review email archives prior to transferring those files to an archival repository.
>
> **Processing:** Provides archivists with the means to arrange and describe email archives.
>
> **Discovery:** Provides the tools for repositories to remotely share a redacted view of email archives with users through a web server discovery environment.
>
> **Delivery:** Enables archival repositories to provide moderated full-text access to unrestricted email archives within a reading room environment.

Requirement	Supporting Functionality	Observations
Support for data transmission	The appraisal module will accept email files directly (from a local file system) and also has the ability connect directly to email servers to download email using IMAP. Other modules rely on outputs (files / directories) from other ePADD modules (i.e. appraisal output is	There are numerous external tools available for moving data. The ability to connect directly to email server is unique and simple if only transporting email content (i.e. no additional

	input to processing module, processing module output is input to discovery module etc.)	content / metadata).
Support for standard formats	Email content can be input in MBOX or by directly connecting to email server (therefore excellent support if only interesting in ingesting email content). It is not possible to input other content (attachments) or Metadata (beyond that contained directly in MBOX format). Email is output to MBOX format. Attachments are NOT extracted separately. Metadata that links correspondents, people, organisations or locations to external authorities (e.g. LC Subject Headings) can be output with URIs that represent the entity by the external authority.	While the format for wrapping metadata appears to be non-standard, the process for assigning the metadata for many descriptive elements (correspondent, location etc.) uses external authorities (linked data) which are well established standards for those specific vocabularies.
Support for appropriate scope of data	ePADD ingests material structured around a particular person who may have more than one email account. It does not appear to offer the wider flexibility of allowing users to enter their own arbitrarily defined 'packets'. It is not possible to input additional metadata or content beyond email / attachments. Processing work can be completed at individual item level (email or attachment) or at various levels of grouping (folder, collection etc.). Additional groupings, such as collections or series, can be added. Scope of outputs can vary as users can select individual emails to include or exclude. Only descriptive metadata can be output (but nothing for rights, sensitivity, processing history etc.) ePADD allows for the re-use or sharing of lexicon files for entity analysis. Lexicon files enable full text searching on a range of different terms, enabling stewards to conduct complex tiered searches.	Metadata can't be input with email content. Metadata can't be output explicitly, but is used in processing so stewards could define workflows that enable them to align to these manually. for example, the cart functionality can be used to select only emails with a certain rights value for output; then repeat for other values, creating an MBOX output file for each metadata value.
Ability to track processing history and provenance	Not available currently.	As noted above, could be some scope for manually outputting data that is grouped around a particular processing 'event' - but no direct support for maintaining, much less merging, processing history.

Support for maintaining the identity and integrity of data	Identifiers are internal (e.g. ePadd message ID) Integration with external authorities such as LC Subject Headings (FAST) ensures consistency and improves accuracy in applying descriptive metadata.	Support for external referencing systems would better enable multi-tool workflows. Linked open data approach for descriptive metadata is unique to ePADD but could be helpful if adopted by other tools.
System Access and Documentation	User documentation available; technical documentation and code available on GitHub.	

4. Key Findings: Analysis of Tools and Email Tools Data Sharing Framework

This section sets out analysis and findings for each of the 'requirements for interoperability' based on our understanding of the capabilities available across all of the tools today. With the exception of some specific integrations (e.g. Archivematica and ArchiveSpace), these tools were not designed to interoperate with each other, and so there are naturally a number of challenges or risks in trying to do that as the tools stand today.

4.1. Current state of data transmission

- Data transmission is, in general, considered out of scope by these tools.
- There is a risk to the chain of custody inherent in any attempt to chain tools together. The primary risk is to metadata that is part of the digital object itself (e.g. created on, created by, modified on, modified by etc.) which can easily be changed or lost as part of 'moving' data from one filesystem to another.
- Many of these tools attempt to minimize this risk internally, e.g., Archivematica, Bitcurator, DArcMail, EAS, all bundle several tools internally and manage data transmission between processing steps.

4.2. Use of standard formats

- Email content for most systems is based on well-established formats, particularly MBOX and EML. So far all systems can input MBOX.
    - EAS outputs only EML and not all tools support this as an input.
- Some systems support only very limited email-specific processing (e.g. Archivematica) and some do not at all (ArchiveSpace) - but as these systems are designed to take in virtually any digital objects this is not a barrier for their more generic processing capabilities
- Identification or referencing metadata is often expected in a 'format' that is non standard in several cases. Message IDs, repository ID, collection ID are often tied to specific external systems (EAS with DRS, DArcMail with CMS).
- PREMIS is the standard used to capture provenance or processing history metadata and rights metadata (for those systems that record this metadata).

- The Library of Congress BagIt standard is a file packaging format used by at at least two of the tools (Archivematica and BitCurator).

4.3. Scope of email data or metadata exchange

- There are no significant barriers to exchanging any particular scope of email content, with the exception that some systems (e.g. ePADD) assume that email is dealt with or managed on an account basis, where an account is the email associated with only one individual. In other words, the user could not input all emails for an entire organisation and process them together at once (while maintaining all individual account level metadata).
- Several tools have limitations on the scope of **metadata** that can be **input** or accepted:
    - EAS, ePadd, DArcMail do not accept any metadata as an input
- Several tools have limitations on the scope of **metadata** that can be **output:**
    - ePADD does not allow for many types of metadata to be output

4.4. Capabilities for recording provenance and/or processing history

- If maintaining a full processing history is necessary, then it may not be feasible to use systems that don't support this (ePADD, DArcMail).

4.5. Capabilities for maintaining identity and integrity of data

Use of unique identifiers:

- Most tools generate unique identifiers for data at various levels of granularity (some for individual email, virtually all for aggregations of some type such as folder, account, collection etc.).
- Most tools do not accept or store 'external' identifiers (i.e. unique IDs created by other systems). This may present challenges when using multiple tools because there are limited ways of ensuring that a particular data item or group of data is correctly identified (for instance, if looking at a particular email in one tool, is there a way of confidently finding and processing the same exact email in another tool).
- Some tools do provide some means of capturing external identifiers (e.g. in Archivematica by providing IDs within a metadata csv file at the point of transfer). However none of the tools appear to support this at the level of individual emails.

Definition of key elements and aggregations:

- Many of the tools allow users to define the elements or aggregations that suit them best. This flexibility is a strength but could lead to some confusion if elements or aggregations are not defined consistently between systems.

- The definition of an Email Account is probably the most significant concern as it appears to be defined differently in different systems. An email account in one tool may appear to be same email account when viewed or processed in another tool, but the risk is that it isn't because the definitions are not consistent. There is also the risk that the data models are not compatible - for instance if one system only allows one email address per account where another allows multiple addresses.

4.6. System access and documentation

- All of the open source systems have publicly available documentation or knowledge resources, however access to developers or subject matter experts may not be publicly available.
- Neither EAS nor DArcMail are currently available beyond their institutions. Both project teams intend to release them with open source licenses, but work is required to do this and make the software available to the community.

5. Opportunities to Improve the Interoperability of Email Tools

Several draft recommendations are suggested below for discussion. At this stage no effort has been made to prioritize these or set out concrete next steps. We have kept the scope of these to areas that we feel address the interoperability of the specific tools assessed in this report.

We have not made any specific recommendations regarding the challenges of transmitting data between systems. While there are some clear risks, as described in the first part of section 4.1 (such as chain of custody and file integrity), we feel that a) these are very broad and apply to all forms of preservation using multiple tools and b) the extent of the problem is not well defined or agreed on; for example, some institutions may not see any problems with data transmission protocols that happen before formal accession. While we feel this area warrants further consideration, that may be outside the scope of concern for this report.

5.1. Enhance tools to support external reference identifiers

At the very least, tools need to be able to accept and maintain external identifiers so that email stewards can keep track (at multiple levels of granularity) what data is being processed throughout a workflow.

In general, email stewards should be able to use the identifiers for individual items, folders or other groupings from one system when exporting data and carrying out further processing in another system.

Ideally external identifiers would also be captured when capturing processing history so that it is possible to clearly track the chain of custody (for example by associating the identifier with the PREMIS agent involved in processing).

5.2. Adopt standard approaches to capturing and respecting rights and sensitivity metadata

Given that email collections often contain content with a variety of different rights, and that there is a wide spectrum of privacy and confidentiality issues that can be involved, email archiving tools should support standard ways of capturing rights or sensitivity metadata.

Many systems already use standards for rights (for instance using PREMIS rights entities); however, there doesn't appear to be an equivalent approach for recording sensitivity or privacy information.

5.3. Establish MBOX as minimum standard for input and output of email content

MBOX is the most widely used standard amongst the tools considered here. EML is also a widely used standard and supported by a majority of email clients. The EAXS standard used in DArcMail may be more comprehensive but has so far not been widely adopted and there are no tools for discovery and access in that format.

We therefore recommend that tool providers consider adding MBOX -- complying with RFC 4155 (Application MBOX Media Type) and RFC 5322 (Internet Message Format) -- as a standard for both input and output (where that doesn't already exist). This doesn't necessarily mean obsoleting use of EML or EAXS, but simply providing additional support to enable maximum interoperability between tools.

5.4. Establish a common exchange standard for packaging email with metadata

A standard for packaging digital content, describing the contents of the package and ensuring integrity of the package using hashes will greatly improve the ability to transfer data safely between systems. The Library of Congress Bagit standard is well-established and is already used by at least two of the tools here (Archivematica and BitCurator).

The BagIt standard may not be enough in itself however. While recommendation 5.3 would ensure that email content can be transferred using the MBOX standard, additional structural and metadata standards may be needed to define minimum expectations for what content or metadata is required, optional or acceptable. For example, to clarify whether it is acceptable to package multiple email accounts together.

5.5. Support capture of processing history

Several tools record processing history using the well established PREMIS standard.

Ideally all tools would provide this capability so that comprehensive processing history can be captured throughout a workflow using multiple tools.

Further consideration should be given to the consolidation of processing history files from different systems, or the ability to manually add processing history (to fill any gaps where a tool does not yet record it automatically).

5.6. Establish standard definition and description of email collections

It isn't clear that the definition of what constitutes an email account (including the relationship with email addresses, or people) is consistent between tools. Establishing a common definition will enable alignment of different data models used and reduce the risk of confusion or mis-identification of email collections at this fundamental level.

With a consistent and standard definition, it will then be possible to develop a common standard for describing email accounts. This would help improve the precision of search and discovery and better enable the exchange of descriptive metadata between tools.

5.7. Make local tools publicly available with an open source license

Tools that are only usable by one institution are not useful to the wider email archiving community. While there are clearly costs to making a tool more widely available and trying to create and maintain an active community around it, we feel there are many benefits that can offset those costs in the long run, including opening up the project to a wider base of developers, testers and potential funders.

Acknowledgements

This project built on the great work started at the Harvard Email Archiving Stewardship Tools (EAST) workshop in March 2016. We would like to thank the original participants and acknowledge the many contributions received since.

In particular we would like to thank the contributors to the Email Data Sharing Framework; Glynn Edwards, Josh Schneider and Peter Chan (Stanford University), Andrea Goethals, Grainne Reilly and Skip Kendall (Harvard University), Sarah Romkey and Justin Simpson (Artefactual Systems Inc.) and Cal Lee (University of North Carolina Chapel Hill).

Numerous reviewers provided helpful contributions and suggestions for this report. We would like to thank Evelyn McLellan, Justin Simpson and Sarah Romkey (Artefactual Systems Inc.), Anthony Moulen, Andrea Goethals and Grainne Reilly (Harvard University), Chris Prom (University of Illinois at Urbana-Champaign), Cal Lee (University of North Carolina Chapel Hill) and Riccardo Ferrante (Smithsonian Institution Archives).

We would like to thank Harvard Library for the opportunity to engage in this work and providing support and direction throughout.

Finally thanks to Wendy Gogel (Harvard University) for contributions on many fronts and providing leadership for the project.

HARVARD
LIBRARY